AI Math Agents

LLMs, Information Systems Biology, Precision Health, and Alzheimer’s Genomics

Melanie Swan,* DIYgenomics, University College London (Research Associate CBT); Takashi Kido, Advanced Comprehensive Research Organization, Teikyo University, Preferred Networks, Inc.; Eric Roland, RedBug AI, LLC; Renato P. dos Santos, Centre for Generative AI in Cognition and Education, Lutheran University of Brazil

Background

Generative AI technologies may be applied to advance science in new ways. LLMs (large language models) are computerized language models generated primarily with transformer neural networks (deep learning methods) having billions of parameters, and pre-trained on large data corpora, e.g. GPT-4 (OpenAI), LMDBA (Google), and LLaMA (Meta AI). Transformer neural networks are an advance which allows a whole data corpus to be processed simultaneously to assess connections between data elements. LLMs are seen as being crucial for genomic analysis given the large scope of data and interrelated parts [1,2].

The important result of LLMs is that they are a linguistic user interface, a language-based access tool, via natural language for human-AI chat (familiar from chatGPT), but more extensively, via formal languages such as programmatic code and mathematics, for further build-out of the computational infrastructure. The entirety of data corpora, not only word-based knowledge bases but also software code and mathematics, are being digitized and mobilized as available easy-to-use tools. Into this trajectory, the current work formulates an AI-based mathematics approach to Alzheimer’s genomics, introducing Math Agents, mathematical embedding, and equation clusters as tools for representing and possibly evaluating genomics, introducing Math Agents, mathematical embedding, and formal languages such as programmatic code and mathematics, for interface, a language-based access tool, via natural language for given the large scope of data and interrelated parts [1,2].

Method

AI-enabled mathematics tools are introduced and demonstrated in the case of Alzheimer’s disease and aging as information systems biology problems. A theoretical model is elaborated, applying multiscalar physics mathematics (elucidating near-far enigmatic correlations in systems) to disease mathematics and whole-human genomic data for two precision medicine patients. Vector embedding as a standard machine learning method is employed with mathematical equations and genomic data as the input.

Results: Alzheimer’s Genomics

Alzheimer’s genomics is a whole-human-genome-based approach to Alzheimer’s disease, developed on GWAS SNPs, EWAS SNPs, eQTL transcripts, transcriptional RNA, and transposon indels. Whereas ApoE profile was the previous means of assessing Alzheimer’s genomics risk, the current understanding includes multiple genomic factors [7]. First, GWAS SNPs may suggest overall risk propensity for the disease. Second, EWAS SNPs indicate which disease genes are actually expressed, confirmed with transcriptome and biomarker assays (Aβ42/40 blood test, CSF). Third, the relation of GWAS-EWAS SNPs on a cis-trans (near-far) correlative basis may be relevant in terms of how different parts of the genome control which genes are expressed. Fourth, transposable elements may be related in activating Alzheimer’s disease through genomic insertion-deletion events produced by viruses or other factors.

The benefit of the mathematical embedding for descriptive and interventional mathematics in Alzheimer’s genomics is that the entirety of a mathscape (set of equations) can be seen in one at-a-glance abstracted and consolidated view. The visualization provides a first-pass view of the mathematics in a paper in the form of equation clusters and mouse-over images. The aggregate view also allows related mathematical ecologies to be compared, and mathematics and data to be investigated in one view as two (ideally corresponding) representations of a system.

References


There is an opportunity to deploy AI-based tools to mobilize mathematics as a high-validation data corpus towards broadly humanity-benefiting use cases in global disease-preventing healthy well-being. Future work could elaborate Math Agents use in an SIR (sustaining, intervening, recovering) model for the societal realization of Precision Health based on two tiers of ongoing circular informational & interventional cycling of the population. Other mathematical discovery use cases of the Math Agent could include synthetic data generation to solve mathematical ecologies, extending equation simplification and math-data-model-fit methods, and employing generative AI with episodic memory (per file distilling-stamping) to assess causal relations in longitudinal personal health dossiers to identify the foundation of pathogenesis. In the short term, genomic variant and eQTL expression data is indicated for practical application to the unresolved challenge of Alzheimer’s disease as the top-five human killer with no survivors.